ARI Research Note 94-25

The Effect of Response Format on Reliability Estimates for Tacit Knowledge Scales

Peter J. Legree
U.S. Army Research Institute

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

Manpower and Personnel Research Division
Zita M. Simutis, Director

July 1994

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Technical review by

Jay Silva

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1994, July	Interim May 93 - Nov 93

4. TITLE AND SUBTITLE

The Effect of Response Format on Reliability Estimates for Tacit Knowledge Scales

5. FUNDING NUMBERS

62785A
790
1211
H01

6. AUTHOR(S)

Legree, Peter J.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8. PERFORMING ORGANIZATION REPORT NUMBER

ARI Research Note 94-25

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

--

11. SUPPLEMENTARY NOTES

--

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

--

13. ABSTRACT (Maximum 200 words)

Most aptitude scales adopt a Forced Choice response format in which answers are scored as either correct or incorrect. Such a scoring procedure is consistent with the nature of the knowledge underlying these aptitude scales because the relevant knowledge domains can usually be used to either support or contradict a specific supposition. Assessing performance with tacit knowledge scales that lack an academic knowledge base often requires the opinions of subject matter experts and responses cannot always be unambiguously scored. Data indicate that an improvement in the reliability of a Tacit Knowledge Scale could be realized by substituting a Likert response format in place of a traditional Forced Choice format; this finding demonstrates the power of the Likert response format to measure individual differences in an uncertain knowledge domain. This research was conducted in support of the development and validation of a Social Intelligence scale.

14. SUBJECT TERMS

Tacit knowledge scale
Likert scale
Practical intelligence reliability
Social intelligence

15. NUMBER OF PAGES

22

16. PRICE CODE

--

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

THE EFFECT OF RESPONSE FORMAT ON RELIABILITY ESTIMATES FOR TACIT KNOWLEDGE SCALES

CONTENTS

List of Tables

THE EFFECT OF RESPONSE FORMAT ON RELIABILITY ESTIMATES FOR TACIT KNOWLEDGE SCALES

Low Fidelity Simulations And Situational Judgment Scales

Situational judgment scales have been developed to measure individual differences in interpersonal skills in a number of areas including telephone sales representative skills (Phillips, 1992), collection agency negotiation skills (Phillips, 1993), administrative and interpersonal skills of educators (Ostroff, 1991), social insight ability (Chapin, 1942), leadership skills of Non-Commissioned Officers (Hanson & Borman, 1992) and managerial skills (Motowidlo, Dunnette, & Carter, 1990). These scales can be descrbed as either situational judgment scales or low fidelity simulations of work sample tasks.

All of these scales are similar in that a Forced Choice format was adopted and each test item is composed of a relatively long problem scenario and a number of actions that might be followed to try to resolve the dilemma. The subject is required to identify the most appropriate action, and sometimes the least appropriate action, based on the problem description, professional knowledge and past experience. A response is scored as either correct or not correct on the basis of agreement with subject matter expert (SME) opinions.

Measuring performance with a low fidelity simulation may appear analogous to testing academic knowledge in that the correct response is externally verified and a response is viewed as either correct or not correct. In addition, the low fidelity simulation methodology allows item level statistics to be computed to identify problematic items and refine the instrument. However, there are several important differences between testing academic knowledge and assessing individual differences in interpersonal skills via a low fidelity simulation.

Item Length One important difference relates to test item length. Academic test items can often be written with relatively short item stems and response choices. In fact, brevity may be recommended to avoid ambiguity in the test, limit the effect of test-wiseness, and maximize the number of items for which data can be collected within some fixed testing period. One implication of this terseness is that the reading requirement of the average item tends to be relatively minimal for many academic knowledge tests.

In contrast, the description of a problem scenario is often lengthy because of the ill-defined and complex nature of these items. Short problem scenarios often cannot support the complexity inherent to the simulated problems.

The response choices also tend to be necessarily longer than those distractors for an academic knowledge test. Thus simulating complex problem scenarios can result in extremely long tests given the number of items for which data are collected. This tendency towards lengthy elaboration is exemplified in the Army Situational Judgment Test, which contains 49 test items and averages one-half page of text per item.

Scoring Another important difference between a standard academic knowledge test and a situational judgment scale involves the procedures used to identify the correct response for a specific item. The scoring key of an academic knowledge test can usually be verified by referencing explicit facts derived from academic theories or listed in reference books. These facts are used to develop a scale that can then be pilot tested. Item level statistics may be computed to identify and either modify or delete problematic test items.

In contrast to the fact-based scoring procedure used for an academic test, scoring a situational judgment scale must often be based on Subject Matter Expert (SME) opinions. Because SME do not always agree, a relatively large number of expert opinions may be required to produce a credible scoring key. For example, Phillips (1992, 1993) required that 75 percent of approximately 20 experts agree that a specific response was "most appropriate", i.e., "correct", in order for that scenario (and response) to be included on a situational judgment scale.

Although 75 percent may appear to be a reasonable agreement criterion, the implication is that for any specific item, up to 25 percent of the experts may disagree as to the most appropriate, i.e., "correct", response alternative. It is relevant to note that with a 75 percent agreement criterion, which is equivalent to an "up to" 25 percent disagreement criterion, the performance of many experts would be far from perfect on a situational judgment scale. In contrast, near perfect performance would often be expected of experts on an academic knowledge test.

It is tempting to conclude that while verification of the correctness of a response alternative is relatively straightforward for academic knowledge, the procedure is quantitatively more complex when applied to the practical knowledge underlying situational judgment scales. However, this interpretation belies the possibility that these two types of knowledge may be qualitatively different. It can be argued that the correctness of an (exam) assertion, given some academic knowledge base (facts and theory), is usually unambiguously dichotomous, i.e., either

correct or not correct.[1] In contrast, situational judgment scales attempt to simulate everyday problem situations but usually cannot present enough information to allow the formulation of unambiguously "correct" solutions. This is not to suggest that if low fidelity simulations could present additional information, then the ambiguity would disappear. This ambiguity partially reflects real-world interpersonal interactions; these are often ambiguous because behavior can be multidetermined and because individuals are dynamic, complex and sometimes disingenuous. It follows that one qualitative difference between academic and everyday interpersonal knowledge is the presence or absence of the certainty that can be attached to the correctness of specific assertions or to the likely result of specific actions given a specific situation or problem.

It is important to recognize that as a general rule, the "correct" response alternative for a low fidelity simulation scenario cannot be guaranteed to always lead to a satisfactory resolution of the simulated problem. Nonetheless, experts will generally agree that some alternatives are much more likely to result in a reasonable solution. A more veridical simulation of expertise for an ill-defined problem situation might require subjects to estimate the relative quality of proposed solutions and compare these estimates to expert ratings. This type of task implicitly recognizes and models this qualitative difference between academic and everyday knowledge.

Likert Scales An alternative to the use of the Forced Choice format is evident in the tacit knowledge scales developed by Wagner and Sternberg (1986). These scales match a single scenario with approximately ten actions and the subjects must rate the appropriateness of all the actions on a Likert scale. The subject ratings can then be transformed to eliminate response bias and a distance is calculated for each item between the transformed subject ratings and the expert (transformed) ratings.

One practical advantage to the Likert format is that one datum is collected for each response alternative as opposed to one or two data per scenario. Therefore this format can be used to collect much more data per unit of text than is possible with the conventional Forced Choice format. For example, a typical scale developed by Wagner and Sternberg (1986) yields ten data points per scenario and requires approximately one page of text. In contrast, the Army Situational Judgment Test yields between

---

[1] It would be possible to create academic test items that do have ambiguous answers. For example, a test could require students to rate the relative clarity of 20 sentences. However, I know of no academic test or scale that utilizes this format. Also, academic test items having unintentionally ambiguous responses are generally dropped after pretest analyses are completed.

two and four data points per page of text.

A second practical advantage to the Likert format is that each response alternative can be converted into a unique item and an interval datum may be computed, i.e., the distance between the subject and expert ratings. This distance quantifies the correctness of the subject's response for a particular item (response alternative) and allows the response to be characterized as varying along the dimension of correctness. In contrast, only dichotomous data are collected with the Forced Choice format. From a practical perspective, it can be expected that the reliability of some existing low fidelity simulations could be substantially improved by incorporating a Likert format in the existing test. This type of modification usually could be implemented with only minor changes to an existing test.

An important conceptual question relating to response format addresses the nature of the task. Most well-designed academic tests, which utilize a Forced Choice format, present a single correct answer per item. The primary purpose of the distractors is to limit the effect of guessing. On this type of scale, the subjects' task can be argued to be primarily an identification task; on this type of task, a knowledgeable subject could respond as soon as a correct response is read.

Unlike a conventional academic multiple choice test, the alternatives for most situational judgment scales are selected to range in correctness (i.e., appropriateness) with several "good" and several "bad" alternatives. When the Forced Choice format is adopted for a situational judgment scale, the task can no longer be considered an identification task because all the alternatives may be somewhat correct without any being optimal. Due to the ambiguity in the problem scenario, none of the actions may be necessarily "best" or even "good." The subject is effectively presented with a complex comparison task that requires the understanding of nuances of vocabulary and meaning, rather than the simple application of the subject's expertise. In addition, the task usually has a substantial memory and reading requirement in order for the lengthy alternatives to be compared.

It could be argued that adapting the Forced Choice format to a situational judgment scale results in a complex vocabulary and memory task; the type of task that typically loads on a general intelligence factor. One possible explanation for the tendency of situational judgment scales to load on general intelligence may be due to this complexity.

By utilizing a Likert format, it may be that the loading of a situational judgment scale on general intelligence will decrease because the importance of these cognitively complex processes, e.g., memory comparison, vocabulary, and reading ability, to task performance will decrease. Individual

differences in interpersonal skills may then become a primary source of variance on that scale, i.e., the divergent validity of the scale relative to general intelligence would increase.

Research Objectives

At present, there is no literature that empirically estimates the effect of utilizing the Likert format on either the reliability of a situational judgment scale or on the empirical relationship of this type of scale to related constructs such as general intelligence. On the basis of classical test theory, i.e., the Spearman-Brown Prophecy formula, it can be hypothesized that an improvement in the reliability of a situational judgment scale would be realized by substituting the Likert format for the forced choice format because the amount of information collected by the scale is greater. One goal of this research is to estimate the extent to which the reliability of an existing situational judgment scale could be improved by utilizing the Likert format.

Another goal of this research is to estimate the empirical relationship between performance on a situational judgement scale and general intelligence. If the construct measured by the scale is independent of the response format, then altering the response format should produce similar correlational estimates between general intelligence and scale performance when corrected for attenuation of reliability. On the other hand, if one effect of the Forced Choice format is to make the scale needlessly complex, then utilizing the Likert format may result in a scale that is less loaded on general intelligence.

This finding would be consistent with the expectation that use of the Likert format could result in better measurement of constructs measured by other situational judgment scales, i.e., telephone sales representative skills (Phillips, 1992), collection agency negotiation skills (Phillips, 1993), administrative and interpersonal skills of educators (Ostroff, 1991), social insight ability (Chapin, 1942), and managerial skills (Motowidlo, Dunnette & Carter, 1990). This research may be relevant to the development and validation of a social intelligence scale because social knowledge tends to be complex and the correctness of an action given a specific situation often cannot be determined, i.e., the correctness of an action lacks certainty. A Likert-based knowledge scale might be a highly efficient format to assess individual differences in social knowledge.

Phase 1

The goal of this phase of the research was to estimate the extent to which the reliability of a situational judgment scale could be improved by utilizing a Likert format. This was accomplished by modifying an existing instrument and testing two groups of subjects with either the Likert or the Forced Choice format.

The Army Situational Judgment Test (SJT) was developed to support Project A research[2] as a test of NCO supervisory ability (Campbell & Zook, 1991). The SJT was selected for modification because it is typical of situational judgment scales in length and response format; in addition, the reported reliability estimates for the scale are in the moderate range (Hanson & Borman, 1992).

Data were collected at the U.S. Air Force Armstrong Data Collection Facility at Lackland AFB. The Lackland subject pool consists of Air Force recruits in their 21st day of basic training. The use of this population necessitated that the SJT content be slightly modified by substituting Air Force specific terms for the Army equivalents. For example, the Air Force rank, Airman, was substituted for the equivalent Army rank, Private.

Materials The SJT consists of 49 problem scenarios with between three and five solutions proposed for each scenario. For Project A, the correct response for each scenario on the SJT had been identified on the basis of SME ratings. The same answer key was used to score the SJT for this research.

The Project A procedure required the subject to read each problem scenario and then to identify the alternative that the subject felt was most appropriate and the alternative that the subject felt was least appropriate. In the current research, the Forced Choice version of the SJT was administered in accordance with instructions and scoring procedures that are essentially identical to those used for Project A.

The Forced Choice version was adapted to the Likert format by appending the following stem to the end of each scenario, "Please rate the appropriateness of the following actions". The instructions for both conditions, which include an example scenario, are contained in Appendix 1. Note that the Likert

---

[2] Project A was a seven year effort designed to justify and improve the procedures used to select and classify Army soldiers. One aspect of Project A was the development and validation of new and existing predictors against job related criteria including the SJT.

response format requires the subject to rate each response alternative, as opposed to selecting the most or least appropriate response.

The subjects were required to rate the appropriateness of each action on an 11 point bipolar scale. The ends of the scale were anchored with the terms "Extremely Inappropriate" and "Extremely Appropriate", the midpoint was labeled "Neither Appropriate Nor Inappropriate". An 11 point scale was used in recognition of the possibility that some scenarios might contain only appropriate or inappropriate alternatives; in such a case, it was felt that a larger interval scale would allow subjects to make finer gradations in their ratings. In addition, Wagner and Sternberg (1986) utilize an 11 point scale and this design was influenced by their research.

Scoring Hanson and Borman (1992) describe a number of ways to score the SJT including: proportion of "most" appropriate hits, proportion of "least" appropriate hits, mean effectiveness SME rating of actions selected as "most" appropriate responses, mean effectiveness SME rating of actions selected as "least" appropriate responses, and the difference between the SME ratings of the "most" and "least" appropriate responses for each scenario. Other situational judgement scales have tended to adopt the simplest of these scoring procedures, i.e., the proportion correct measures.

The two proportion measures were calculated by defining the response alternatives that were rated highest and lowest by the SME as the correct "most" appropriate and "least" appropriate response for each scenario. Individual difference scores were calculated as the proportion of correct responses for the two dimensions.

The mean effectiveness rating scoring procedures weighted the "most" and "least" appropriate response for each scenario by the mean SME ratings for those responses. Thus if a subject selected a response alternative with a mean SME rating of 5.27 as the "most" appropriate response, then a value of 5.27 would be assigned for that item. Accordingly, better performance is indicated by higher scores for the "most" appropriate responses and by lower scores for the "least" appropriate responses. The difference measure was calculated by averaging (across scenarios) the difference in the weightings associated with the "most" appropriate and "least" appropriate responses. In this study, all five procedures were used.

The procedure used to score the Likert version of the SJT is dissimilar from any typically used to calculate individual

differences on ability scales[3]. The procedure produces interval data for each item as a function of the distance between the subject's rating and the expert rating for that response alternative. The average distance across items is then computed to estimate individual differences in performance on the task. However, several transformations of the data are required to eliminate response bias.

Response bias is an important issue because the scoring procedure is intended to quantify individual differences in the ability to estimate the relative appropriateness of alternate solutions given a specific problem scenario. If ignored, response bias could have a dramatic effect for subjects who use only part of the rating scale. For example, if the ratings of a particular subject were biased towards the "Inappropriate" segment of the scale, then the distances calculated for all but the most inappropriate alternatives would be overestimated.

To resolve the response bias problem, the ratings produced by each subject were transformed to yield standard scores with a mean of 0.0 and a standard deviation of 1.0. A similar transformation was conducted on the expert ratings of the effectiveness of the alternatives described for the scenarios. These SME ratings had been collected as part of Project A. A distance was then calculated for each item as the square of the difference between the transformed expert and subject ratings. Individual differences were computed as the mean item distance for each subject. Using this procedure, better performance is indicated by lower values.

Subjects Forty-eight male Air Force recruits in their 21st day of basic training at Lackland AFB participated in this study. Twenty-four subjects were assigned to each group.

Procedure Data were collected after breakfast over a two week period between 7:00 and 9:00 AM. Subjects were alternately assigned to a condition, i.e., Likert versus Forced Choice. The subjects were seated in a classroom and instructed to follow the instructions described in the SJT testbook. The subjects were tested in groups of up to 20 subjects and were told to wait at their desks until the session was completed.

Results Reliability estimates were calculated for the two versions of the scale and indicate a substantial increase in the

---

[3] Some important terminology changes must be noted. A Likert alternative corresponds in content to a Forced Choice response alternative, but data are collected for all Likert alternatives while most response alternatives are distractors within the Forced Choice format. The terms "distractor" and "p-value" are meaningless from a distance perspective.

reliability estimate of the Likert format relative to the Forced Choice format. Table 1 contains estimates of the reliability, the mean performance of the subjects, and the standard deviation of performance by scoring procedure.

One advantage to the Likert format is that it is possible to eliminate items at the alternative level. To demonstrate this advantage, total and item score were correlated across the 202 Likert items. Fifty-seven items with negative full scale correlations were eliminated from the scale and the reliability of the new scale was estimated at .84 (Refer to Table 1).

A similar procedure was followed for the most reliable scoring method that is associated with the Forced Choice format, i.e., the Difference weighting procedure. Seven scenarios with negative total score correlations were deleted from the scale, with the result that the reliability of the revised scale increased to .65 (Refer to Table 1).

Table 1. Test Means, Standard Deviations, and Internal Consistencies.

Scoring Procedure	Mean	SD	Reliability
Forced Choice Format			
Most Proportion Correct	.46	.07	.27
Least Proportion Correct	.45	.08	.31
Most Weighting	4.79	.21	.37
Least Weighting	3.38	.17	.26
Difference Weighting	1.41	.36	.51
Likert Format			
Alternative Level	1.19	.18	.62
Refined Scales			
Forced Choice (42 scenarios)			.65
Likert (145 items)			.84

One question revolving around the use of SME ratings to reference subject performance relates to the relationship between mean SME and the Air Force subject ratings across the 202 Likert alternatives. Agreement in mean ratings was assessed by correlating the two sets of mean ratings (r=.72, p<.01).

Discussion The reliability data indicate that the Likert format, when utilized in place of the Forced Choice format, results in much more reliable individual difference estimates. In one sense, this is to be expected because the number of data points is dramatically increased. Increasing the amount of collected data should, according to the Spearman-Brown prophecy formula, result in a much more reliable scale. This demonstration would be trivial if the increase in reliability was simply due to increasing the length of the scale by adding additional items; but this was not the case.

The reliability data are important specifically because the additional data were collected with only minimal differences in the length of the scale, i.e., the amount of information that was presented to the subjects. In fact, the complexity of the scale arguably decreased because the Likert task does not require the multiple comparisons associated with the forced choice format.

The improvement in reliability is also notable because this issue has not been empirically addressed and previously reported. This basic methodology could be applied to improve the psychometric properties of a number of existing scales, i.e., those listed in the introduction of this report. One might wonder what result a factor analysis of such batteries would produce. Future research could easily address this question.

The reason that this format has not been utilized more fully in the past is not clear, the scoring procedure is not conceptually complex. One obvious limitation to the Likert scoring procedure is that the transformations and distance calculations require a substantial amount of computing time per subject. Although this procedure could be performed manually, it seems unlikely that much research with this format would be conducted by individuals who are not familiar with a programming language. Possibly these computational requirements have discouraged the construction of ability or aptitude scales based on a Likert format.

One might speculate that the Likert format could be used to develop a tacit knowledge scale oriented towards knowledge domains that are associated with specific personality traits. For example, individual differences in emotional stability might correlate with knowledge of either the relative effectiveness of strategies that can be used to attenuate feelings of emotional distress or the extent to which specific situations are likely to result in emotional distress. Likewise, individuals who are high

in assertiveness or dominance would be expected to know more about being assertive or dominant.

The major advantage to a personality scale based on tacit knowledge over existing personality inventories is that faking is not an issue on such a scale. This is because the tacit knowledge format explicitly requires subjects to estimate the objective appropriateness of various actions, i.e., the most correct responses are also the most socially desirable. Existing personality inventories usually require subjects to describe their personalities through agreement with various statements and many of these statements tend to be loaded in social desirability. As a result, existing personality inventories are often highly fakable, e.g., instructing subjects to "fake good" on a personality inventory resulted in a 1.7 standard deviation unit increase on the composites (Young, White & Oppler, 1991; Young, White & Oppler, 1992). The important point is that a tacit knowledge scale that correlates with a personality trait could be used to predict performance and support personnel selection and classification.

One difference between this research and Project A is that the Project A subjects were soldiers with substantial military experience, while the Air Force recruits were in their 21st day of basic training. This difference could be important because the SJT contains scenarios that require the subjects to assume the role of a military supervisor confronted with a variety of personnel problems. If correct performance was based on explicit military doctrine, then it would be expected that the performance of the Project A subjects would be substantially better than the Armstrong Air Force recruits.

However, the development of the SJT did not utilize explicit military knowledge or doctrine. Instead Army supervisors were contacted to identify knowledge that can be considered primarily tacit, which is why Army SME ratings were needed to identify the correct responses. Many of the simulated personnel problems may tap interpersonal experiences that transcend military or civilian settings. According to this argument, only small differences in performance should be apparent in the comparison of the Army and Air Force samples.

If military experience leads to a general improvement in performance on the SJT, there should be a substantial difference in mean performance between the Army and Air Force samples. Descriptive statistics obtained from Hanson and Borman (1992) and computed for the Force Choice condition in this research are reported in Table 2. Effect sizes were calculated in accordance with the approach described by Bloom (1984) and are reported in Table 2. The Army variance estimates were used because these are based on a much larger sample size. The effect size estimates are consistent with the interpretation that performance on the

SJT is not highly influenced by military experience and may tap general interpersonal knowledge. Three of the five comparisons, including the most reliable scale, actually favor the Air Force subjects.

Table 2. Comparison of Army and Air Force Levels of Performance.

Scoring Procedure	Army Mean (SD)	Army r_{xx}	Air Force Mean (SD)	Air Force r_{xx}	Effect Size[1]
Most Proportion Correct	.47 (.12)	.60	.46 (.07)	.27	.08
Least Proportion Correct	.42 (.11)	.57	.45 (.08)	.31	-.27[2]
Most Weighting	4.91 (.34)	.68	4.79 (.21)	.37	.35
Least Weighting[3]	3.54 (.31)	.68	3.38 (.17)	.26	.52[2]
Difference Weigthing	1.36 (.61)	.75	1.41 (.36)	.51	-.08[2]

[1] Calculated in accordance with Bloom (1984) with the Army SD as the reference value.
[2] The difference favors the Air Force sample.
[3] Low scores indicate better performance.

There were, however, substantial differences in the reliability and the variance estimates associated with the Forced Choice format reported and estimated on the basis of either the Project A or the Air Force data. In general, the Air Force recruits are highly selected in that the Aptitude Area scores required to enter Air Force specialties tend to be high. If performance on the SJT is heavily g-loaded, then the variance in performance of the Air Force recruits would be more attenuated than the Army sample due to implicit restriction of range. This attenuation would be consistent with the lower variance and reliability values estimated for the Forced Choice format based on the Air Force sample then was estimated based on the Army data. In any event, the Air Force and Army reliability estimates are consistent with the differences in estimated variance.

Future Research

The first phase data suggest that a substantial improvement in the reliability of the SJT could be realized by substituting the Likert format for the Forced Choice format. However, the reliability estimates were based on small sample sizes (24 per group) and inferential tests were not applied. It is possible that the observed differences were due to sampling error and therefore a much larger sample size is intended in Phase 2 to compute more stable reliability estimates.

The first phase of this project also did not address the relationship between performance on the SJT and measures of general cognitive ability. The second phase of this research intended to address this question.

Data are being collected at Lackland AFB. A total of 400 subjects will be tested on the SJT with either the Likert or the Forced Choice format. These sample sizes will allow the above questions to be addressed and the analyses will be reported as they become available.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 4-16.

Campbell, J. P., & Zook, L. M. (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A. (ARI Research Report 1597). Alexandria, VA: U. S. Army Reasearch Institute for the Behaviiora. and Social Sciences. (AD A242 921)

Chapin, F. S. (1942). Preliminary standardization of a Social Insight Scale. American Sociological Review, 7, 214-228.

Hanson, M. A., & Borman, W. C. (1992). Development and construct validation of the situational judgment test (SJT). PDRI Institute Report #230. Minneapolis, MN: Personnel Decisions Research Institute Incorporated.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selction procedure: The low-fidelity simulation. Journal of Applied Psychology, 75, 640-647.

Ostroff, C. (1991). Training effectiveness measures and scoring schemes: A comparison. Personnel Psychology, 44, 353-374.

Phillips, J. F. (1992). Predicting sales skils. Journal of Business and Psychology, 7, 151-160.

Phillips, J. F. (1993). Predicting negotiation skills. Journal of Business and Psychology, 7, 403-411.

Wagner, R. K. (1985). Tacit knowledge in everyday intelligent behavior (Doctoral dissertation, Yale University, 1985). Dissertation Abstracts International, 46, 4049.

Wagner, R. K. (1986). The search for interterrestrial intelligence. In R. Sternberg & R. Wagner (Eds.), Practical intelligence: Nature and origins of competence in the everyday world (pp. 361-378). New York, NY: Cambridge University Press.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. Journal of Personality and Social Psychology, 52, 1236-1247.

Wagner, R. K., & Sternberg, R. J. (1986). Tacit knowledge and intelligence in everyday world. In R. Sternberg & R. Wagner (Eds.), Practical intelligence: Nature and origins of competence in the everyday world (pp. 51-83). New York, NY: Cambridge University Press.

Young, M. C., White, L. A., & Oppler, S. H. (1992, October). Effects of coaching on validity of a self-report temperament measure. Paper presented at the meeting of the Miliary Testing Association, San Diego, CA.

Young, M. C., White, L. A., & Oppler, S. H. (1991, October). Coaching effects on the assessment of background and life experiences (ABLE). Paper presented at the meeting of the Military Testing Association, San Antonio, TX.

Appendix A

Instructions for the Two Situational Judgment Test Conditions

Page A-2 contains the instructions for the Likert version of the SJT; an example of a scenario and completed ratings is embedded in the instructions. Page A-3 contains the instructions and the example used for the Forced Choice version of the SJT. The example was modified for the Air Force subjects by replacing the Army term, Platoon, with the equivalent Air Force term, Flight.

SITUATIONAL JUDGEMENT TEST

In this booklet, you will be presented with a series of supervisory situations. These are situations in which a first line supervisor might find him/herself. After each situation several possible responses to that situation are listed. To insure realistic scenarios, the situations and responses are based on the experiences and statements of senior NCOs.

Your task is to read each situation and the responses listed. Then rate the appropriateness of each of the actions on the 11 point scale. Be sure to rate all the actions.

Below is an example of an item that has been completed properly.

---

1 2 3 4 5 6 7 8 9 10 11

Extremely Inappropriate — **Neither Appropriate Nor Inappropriate** — **Extremely Appropriate**

You are a Work Center NCOIC. Over the past several months you have noticed that one of the other Work Center NCOICs in your Flight hasn't been conducting his Common Task Training (CTT) correctly. Although this hasn't seemed to affect the Flight yet, it looks like the Flight's marks for CTT will go down if he continues to conduct CTT training incorrectly. How appropriate are the following actions.

2 a. Do nothing since performance hasn't yet been affected.

7 b. Have the Work Center NCOIC meeting and tell the Work Center NCOIC who has been conducting training improperly that you have noticed some problems with the way he is training his troops.

8 c. Tell your Flight sergeant about the problem.

10 d. Privately pull the Work Center NCOIC aside, inform him of the problem, and offer to work with him if he doesn't know the proper CTT training procedure.

You may not agree with the ratings for this item, but this example shows you how these items should be completed.

Be sure to rate each item on the 11 point scale and be sure to use the entire scale.

SITUATIONAL JUDGEMENT TEST

In this booklet, you will be presented with a series of supervisory situations. These are situations in which a first line supervisor might find him/herself. After each situation several possible responses to that situation are listed. To insure realistic scenarios, the situations and responses are based on the experiences and statements of senior NCOs.

Read each situation and the responses listed. Then decide which of these possible responses would be the most effective. Place an "M" in the box next to the most effective response.

Next decide which of these possible responses is the least effective. Place an "L" in the box next to the least effective response. The boxes in front of the remaining response alternatives should be left blank.

Below is an example of an item which has been completed properly.

You are a Work Center NCOIC. Over the past several months you have noticed that one of the other Work Center NCOICs in your Flight hasn't been conducting his Common Task Training (CTT) correctly. Although this hasn't seemed to affect the Flight yet, it looks like the Flight's marks for CTT will go down if he continues to conduct CTT training incorrectly. What should you do?

__L__ a. Do nothing since performance hasn't yet been affected.

___ b. Have the Work Center NCOIC meeting and tell the Work Center NCOIC who has been conducting training improperly that you have noticed some problems with the way he is training his troops.

___ c. Tell your Flight sergeant about the problem.

__M__ d. Privately pull the Work Center NCOIC aside, inform him of the problem, and offer to work with him if he doesn't know the proper CTT training procedure.

You may not agree with the placement of the "M" and the "L" for this item, but this example shows you how these items should be completed.

In summary, for each item you will place an "M" for Most effective next to one response alternative, and an "L" for Least effective next to another response alternative. The boxes in front of the rest of the response alternatives will be left blank. Please use only one "M" and only one "L" per item.